12th International Conference on Information Fusion
Seattle, WA, USA, July 6-9, 2009

# Diffuse Prior Monotonic Likelihood Ratio Test for Evaluation of Fused Image Quality Metrics *

**Chuanming Wei**
ECE Dept.
Lehigh University
Bethlehem, PA 18015
chw207@lehigh.edu

**Lance M. Kaplan**
Army Research Lab.
Adelphi, MD 20783
lance.m.kaplan@us.army.mil

**Stephen D. Burks**
NVESD
Fort Belvoir, VA 22060
stephen.burks1@us.army.mil

**Rick S. Blum**
ECE Dept.
Lehigh University
Bethlehem, PA 18015
rblum@ece.lehigh.edu

**Abstract** − *This paper introduces a novel method to score how well proposed fused image quality measures (FIQMs) indicate the effectiveness of humans to detect targets of interest in fused imagery. The human detection performance is measured via human perception experiments. A good FIQM should relate to perception results in a monotonic fashion. The new method, the diffuse prior monotonic likelihood ratio (DPMLR) test, compares the $H_1$ hypothesis that the intrinsic human detection performance is related to the FIQM via a monotonic function to the null hypothesis that the detection and image quality relationship is random. The paper discusses many interesting properties of the DPMLR and demonstrates the effectiveness of the DPMLR test via Monte Carlo Simulations. Finally, the DPMLR is used to score FIQMs over 35 scenes implementing various image fusion algorithms.*

**Keywords:** Image fusion, fused image quality measures, hypothesis test, monotonic correlation.

## 1 Introduction

In recent years, image fusion has been attracting a large amount of attention in a wide variety of applications such as concealed weapon detection [1], remote sensing [2], intelligent robots [3], medical diagnosis [4], and military surveillance [5]. Image fusion refers to generating a fused image in which each pixel is determined from a set of pixels in each source image. The fused image should contain a better view of the scene than do any of the source images, thus improving computer or human interpretation. The interested reader is referred to Chapter 1 of [6] for a survey of various image fusion algorithms developed in past years.

Measuring the performance of image fusion algorithms is an extremely important task which has received past study [7–21]. The performance of image fusion algorithms is primarily assessed by perceptual evaluation in the form of subjective human tests [13]. In these tests, human observers are asked to view a series of fused images and rate them. Although the subjective tests are typically accurate if performed correctly, they are inconvenient, expensive and time consuming. Hence, we desire an objective performance measure that can accurately predict human perception. Note that here we refer to the metrics and features proposed for evaluating the quality of the fused images as fused image quality measures (FIQMs). In the literature, there are three broad classes of FIQMs. The first class requires a reference fused image (or the ground truth image), while the others don't. In some special cases (for instance, the multi-focus image fusion [8]), it is possible to generate such a reference image. Once the ground truth image is given, we can use existing quality metrics such as the mean square error and the peak signal to noise ratio to compare the experimental fused results with the reference. However, in many applications generating the ideal fused image is usually very difficult. For this reason we do not consider FIQMs which requre reference image in this paper. Another class of FIQMs introduced recently have received a lot of attention [9–12]. These measures, see [14], consider the sum

*Research was sponsored by the Army Research Laboratory under grant W911NF-06-2-0020 and the U.S. Army Research Office under grant W911NF-08-1-0449, and by a grant from the Commonwealth of Pennsylvania, Department of Community and Economic Development. The views and conclusions contained in this document are those of the authors and should not be interpreted as the official policies, either express or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **JUL 2009** | | **06-07-2009 to 09-07-2009** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **Diffuse Prior Monotonic Likelihood Ration Test for Evaluation of Fused Image Quality Metrics** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **ECE Department,Lehigh University, ,Bethlehem,PA** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES
**See also ADM002299. Presented at the International Conference on Information Fusion (12th) (Fusion 2009). Held in Seattle, Washington, on 6-9 July 2009. U.S. Government or Federal Rights License.**

14. ABSTRACT
**This paper introduces a novel method to score how well proposed fused image quality measures (FIQMs) indicate the effectiveness of humans to detect targets of interest in fused imagery. The human de- tection performance is measured via human perception experiments. A good FIQM should relate to perception results in a monotonic fashion. The new method, the diffuse prior monotonic likelihood ratio (DPMLR) test compares the H1 hypothesis that the intrinsic human de- tection performance is related to the FIQM via a mono- tonic function to the null hypothesis that the detection and image quality relationship is random. The paper discusses many interesting properties of the DPMLR and demonstrates the effectiveness of the DPMLR test via Monte Carlo Simulations. Finally, the DPMLR is used to score FIQMs over 35 scenes implementing var- ious image fusion algorithms.**

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT **unclassified** | b. ABSTRACT **unclassified** | c. THIS PAGE **unclassified** | **Public Release** | **8** | |

of correlations between each source image and the fused image, which provides a measurement of the amount of information transferred from the source images to the fused image. The third class of FIQMs tries to extract the salient features, such as the structure, texture, contrast and edge information, directly from the fused image without regard to the source images [17–21]. Comparisons between existing FIQMs have been lacking.

Quantitatively evaluating the image fusion performance is a complicated issue because of the lack of a complete understanding of the human visual system (HVS), and because of the variety of image fusion applications [15]. We expect that the FIQM should be task specific, and the best measure changes from task to task. Given an image fusion application and many kinds of proposed FIQMs, we are interested in which quality measure describes the image fusion performance better. Clearly, a good FIQM must be related to how a human would judge the quality of the fused image in a monotonic fashion. Therefore, a statistic which quantifies how well different FIQMs are consistent to actual human performance is necessary, which is the focus of this paper.

In [16], Pearson (or linear) correlation and root mean squared error (RMSE) are used to score potential FIQMs. The Pearson correlation is the most common method to determine whether or not the input and output sequences are related. It quantifies how well a straight line fits mapping between the input and output sequences. Unfortunately, when the relationship between the quality measure and the human performance is nonlinear, the value of Pearson correlation can be small despite the fact that the sequences are still monotonically related. In essence, a proper statistic needs to determine if the ordering of quality measures preserve the ordering of the corresponding human performance measures. A nonlinear correlation coefficient referred to as the monotonic correlation (MC) has be proposed in [17]. The MC is more general than the Pearson correlation and exploits the monotonic regression between the quality measures and the human observations. However, it assumes that the perception error is Gaussian, which should be fine for a large number of observers. In this paper, we take a different approach. We focus on cases where the fused image is to be used for object detection. Performance is measured by the probability that a human observer can correctly detect certain objects of interest in the fused image. We introduce a new monotonic statistic for the object detection task where the underlying perception results should follow a binomial distribution and the number of observers is small.

The paper is organized as follows. Section 2 presents the perception model and introduces the new monotonic statistic. Section 3 demonstrates the effectiveness of the new statistic via Monte Carlo simulations. The statistic is used to score potential FIQMs against actual perception results for fused image interpretation in Section 4. Finally, Section 5 provides some concluding remarks.

# 2 Monotonic Statistic

## 2.1 Data Models

This paper considers the detection task so that the performance of image fusion algorithms is the probability that a human observer can correctly detect certain objects of interest in the fused image. A scene is a realization of $F$ source images, and $N$ fused images are generated from these $F$ images via $N$ different algorithms. The existence (or lack) of a monotonic relationship between measured human performance and computed FIQMs can be inferred over $S$ scenes. To this end, this subsection provides the data models that enable this inference.

For a given scene, let the $N \times 1$ vector $\mathbf{p}$ denote the actual performance for all fusion methods, where $p_i$ is the object detection probability associated to the $i$-th fused image. The value of $\mathbf{p}$ is unobservable. It can only be inferred via perception experiments that measure $\mathbf{y}$ where $y_i$ is the number of observers that correctly detect the targets in the $i$-th fusion image. We use $o_i$ to represent the number of observers that participate in the detection experiment for the $i$-th fusion image. It is reasonable to model $\mathbf{y}$ as a random vector whose elements are statistically independent where $y_i$ is drawn from a binomial distribution with parameters $o_i$ and $p_i$, i.e.,

$$\mathbf{y} \sim f(\mathbf{y}|\mathbf{o},\mathbf{p}) = \prod_{i=1}^{N} \binom{o_i}{y_i} p_i^{y_i}(1-p_i)^{o-y_i}. \qquad (1)$$

Here we collect $(o_1 = \cdots = o_N)$ in an $N \times 1$ vector $\mathbf{o}$ for convenience.

Let a given FIQM evaluated over the $N$ images be denoted as $\mathbf{x}$. The measure value $x_i$ is a deterministic function of the $i$-th fused image and the $F$ source images. However, over the ensemble of all possible scenes, the value of $x_i$ can be viewed as a statistical quantity. The proposed monotonic hypothesis test evaluates how well a FIQM monotonically relates to human object detection performance. Under the monotonic hypothesis, there is a monotonic function that maps the measure value $x_i$ associated to the $i$-th fusion method to the detection probability $p_i$, i.e.,

$$p_i = g(x_i), \qquad (2)$$

where $g(x)$ is a monotonic increasing or decreasing function of $x$. For notational convenience, we index the $N$ image fusion algorithms in ascending order of the corresponding measure values, i.e., $x_1 \leqslant x_2 \leqslant ... \leqslant x_N$. Thus, we consider two alternative $H_1$ hypotheses: $H_\uparrow$ for ascending $p_i$'s and $H_\downarrow$ for descending $p_i$'s. On the

other hand, the null hypothesis is that over the ensemble of possible fused imagery, the $x_i$'s are *i.i.d.* samples. Thus, the $p_i$'s are in random order where the probability of any permutation of the order is equal.

A given scene is a realization from the ensemble of possible source images. Therefore, we can model the detection probabilities as being drawn from a random distribution. We use an uninformative (or diffuse) prior for the hypotheses. For the $H_\uparrow$, $H_\downarrow$, and $H_0$ hypotheses, $\mathbf{p}$ is uniformly distributed over

$$\begin{aligned} \mathcal{P}_\uparrow &= \{\mathbf{p} : 0 \leqslant p_1 \leqslant \ldots \leqslant p_N \leqslant 1\}, \\ \mathcal{P}_\downarrow &= \{\mathbf{p} : 1 \geqslant p_1 \geqslant \ldots \geqslant p_N \geqslant 0\}, \quad \text{and} \\ \mathcal{P}_0 &= \{\mathbf{p} : 0 \leqslant p_1, \ldots, p_N \leqslant 1\}, \end{aligned} \tag{3}$$

respectively. Over all hypotheses, we model the $\mathbf{p}_s$'s for each scene as statistically independent of each other.

## 2.2 Diffuse Prior Monotonic Likelihood Ratio Test

The proposed monotonic statistic leads to a hypothesis test that is designed to work for a small number of observers. It exploits the binomial distribution of the perception results by considering the likelihoods for each of the hypotheses. The ascending and descending likelihoods are given by (1). Because the ordering of the elements of $\mathbf{y}$ and $\mathbf{o}$ are random for the null hypothesis, the likelihood of $\mathbf{p}$ is not dependent on the orderings of the observations. In short, the hypothesis test distinguishes between the three following likelihoods

$$\begin{aligned} l(H_\uparrow|\mathbf{y},\mathbf{o},\mathbf{p}) &= f(\mathbf{y}|\mathbf{o},\mathbf{p}) && \text{for } \mathbf{p} \in \mathcal{P}_\uparrow, \\ l(H_\downarrow|\mathbf{y},\mathbf{o},\mathbf{p}) &= f(\mathbf{y}|\mathbf{o},\mathbf{p}) && \text{for } \mathbf{p} \in \mathcal{P}_\downarrow, \\ l(H_0|\mathbf{y},\mathbf{o},\mathbf{p}) &= \tfrac{1}{N!}\textstyle\sum_{j=1}^{N!} f(P_j\mathbf{y}|P_j\mathbf{o},\mathbf{p}) && \text{for } \mathbf{p} \in \mathcal{P}_0, \end{aligned} \tag{4}$$

where $P_j$ is one of the $N!$ possible $N \times N$ permutation matrices. The resulting hypothesis tests are not simple tests because $\mathbf{p}$ is not observable. One could resort to a generalized likelihood ratio, but the resulting test is not a universally most powerful test. Alternatively, we use the uninformative model for $\mathbf{p}$ as given in Section 2.1 and calculate the expected likelihood via

$$\tilde{l}(H_i|\mathbf{y},\mathbf{o}) = \int_{\mathcal{P}_i} l(H_i|\mathbf{y},\mathbf{o},\mathbf{p}) f(\mathbf{p}|H_i)\, d\mathbf{p}, \tag{5}$$

where $i \in \{\uparrow, \downarrow, 0\}$ and $f(\mathbf{p}|H_i)$ is the uniform probability density function over $\mathcal{P}_i$:

$$f(\mathbf{p}|H_i) = \begin{cases} \left(\int_{\mathcal{P}_i} d\mathbf{p}\right)^{-1} & \text{for } \mathbf{p} \in \mathcal{P}_i, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

It is easy to see that

$$\int_{\mathcal{P}_0} d\mathbf{p} = 1, \quad \text{and} \quad \int_{\mathcal{P}_\uparrow} d\mathbf{p} = \int_{\mathcal{P}_\downarrow} d\mathbf{p} = \frac{1}{N!}. \tag{7}$$

The diffuse ascending and descending likelihood ratios to test the $H_\uparrow$ and $H_\downarrow$ hypotheses, respectively, are given by:

$$\lambda_N^\uparrow(\mathbf{y},\mathbf{o}) = \frac{\tilde{l}(H_\uparrow|\mathbf{y},\mathbf{o})}{\tilde{l}(H_0|\mathbf{y},\mathbf{o})}, \tag{8}$$

$$\lambda_N^\downarrow(\mathbf{y},\mathbf{o}) = \frac{\tilde{l}(H_\downarrow,\mathbf{y},\mathbf{o})}{\tilde{l}(H_0|\mathbf{y},\mathbf{o})}. \tag{9}$$

For multiple scenes, the overall likelihood ratios are the product of the single scene likelihoods due to the uninformative model of $\mathbf{p}$ given in Section 2.1. The likelihood ratio for the monotonic relationship is

$$\Lambda_N = \max\left[\prod_{s=1}^{S} \lambda_N^\uparrow(\mathbf{y}_s,\mathbf{o}_s), \prod_{s=1}^{S} \lambda_N^\downarrow(\mathbf{y}_s,\mathbf{o}_s)\right], \tag{10}$$

where $\mathbf{y}_s$ and $\mathbf{o}_s$ are the number of correct detections and observations for the $s$-th scene, respectively. Unless it is required, the scene index is implicit for the sake of notational brevity. We refer to $\Lambda_N$ as the diffuse prior monotonic likelihood ratio (DPMLR). When $\Lambda_N > 1$ the evidence in support of the monotonic hypothesis is greater than that of the null hypothesis where the FIQM behaves as noise with respect to human performance. As $\Lambda_N$ increases, so does the evidence that the FIQM under test is actually a good measure. The DPMLR test is simply accepting the monotonic hypothesis if the DPMLR exceeds a given threshold value. Usually, the threshold is greater than one.

## 2.3 Recursive Computation

To our knowledge, a closed form expression for (8) and (9) does not exist. It is possible to calculate the diffuse likelihood ratios numerically. However, due to the multiple integration involved in the expression, the calculation requires large computational cost, especially when $N$ and the $o_i$'s are large. This subsection provides a recursion to calculate these diffuse likelihood ratios.

The diffuse likelihood for $H_0$ can be simply expressed as:

$$\tilde{l}(H_0|\mathbf{y},\mathbf{o}) = \prod_{i=1}^{N} \binom{o_i}{y_i} \beta(y_i + 1, o_i - y_i + 1) \tag{11}$$

where

$$\beta(a,b) = \int_0^1 z^{a-1}(1-z)^{b-1}\, dz \tag{12}$$

is the Beta function.

Substituting equations (1), (5) and (11) into (8), the ascending diffuse likelihood ratio can be expressed as:

$$\lambda_N^\uparrow(\mathbf{y},\mathbf{o}) = \frac{N! \int_{\mathcal{P}_\uparrow} h(p_N; y_N, o_N) \ldots h(p_1; y_1, o_1) d\mathbf{p}}{\prod_{i=1}^{N} \beta(y_i + 1, o_i - y_i + 1)}, \tag{13}$$

where

$$h(p; y, o) = p^y (1-p)^{o-y}. \tag{14}$$

By considering the power series expansion of the regularized incomplete Beta function, (13) can be simplified. Specifically, the regularized incomplete Beta function is defined as

$$I(y;a,b) = \frac{\int_0^y z^{a-1}(1-z)^{b-1}\,dz}{\beta(a,b)}, \tag{15}$$

and the power series expansion for $I(y;a,b)$ is

$$\begin{aligned} &I(y;a,b) = \\ &= \frac{1}{a+b}\sum_{j=a}^{a+b-1}\frac{1}{\beta(j+1,a+b-j)}y^j(1-y)^{a+b-1-j}. \end{aligned} \tag{16}$$

Now (13) can be simplified to:

$$\begin{aligned} &\lambda_N^{\uparrow}(\mathbf{y},\mathbf{o}) = \\ &= \frac{N!}{o_1+2}\sum_{j=y_1+1}^{o_1+1}\frac{\beta(j+y_2+1,o_1+o_2+2-y_2-j)}{\beta(j+1,o_1+2-j)\beta(y_2+1,o_2-y_2+1)}\cdot \\ &\frac{\int_0^1\ldots\int_0^{p_3}h(p_N;y_N,o_N)\ldots h(p_2;j+y_2,o_1+o_2+1)\,dp_2\ldots dp_N}{\prod_{i=3}^N\beta(y_i+1,o_i-y_i+1)\beta(j+y_2+1,o_1+o_2+2-y_2-j)} \\ &= \frac{N!}{o_1+2}\sum_{j=y_1+1}^{o_1+1}\frac{\beta(j+y_2+1,o_1+o_2+2-y_2-j)}{\beta(j+1,o_1+2-j)\beta(y_2+1,o_2-y_2+1)}\cdot \\ &\lambda_{N-1}^{\uparrow}\left([j+y_2,y_3,\ldots,y_N]',[o_1+o_2+1,o_3,\ldots,o_N]'\right). \end{aligned} \tag{17}$$

Also note that by definition,

$$\lambda_1^{\uparrow}(y_1,o_1) = 1. \tag{18}$$

Thus the ascending diffuse likelihood ratio can be computed numerically via the recursion defined in (17) and (18). A similar recursion can compute the descending diffuse likelihood ratio. Alternatively, one can use (17) and (18) and exploit the fact that

$$\begin{aligned} &\lambda_N^{\downarrow}\left([y_1,\ldots,y_N]',[o_1,\ldots,o_N]'\right) = \\ &\quad = \lambda_N^{\uparrow}\left([o_1-y_1,\ldots,o_N-y_N]',[o_1,\ldots,o_N]'\right). \end{aligned} \tag{19}$$

The symmetric relationship in (19) can be proved by a simple change of variables in (13).

### 2.4 Properties

For the common case that the number of observers in the perception experiment are consistent over the different fused imagery, i.e., $o_i = o$ for $1 \leqslant i \leqslant N$ ($\mathbf{o} = o\mathbf{1}$), the diffuse likelihood ratios have some interesting properties:

1. If $y_1 = y_2 = \ldots = y_N$, then $\lambda_N^{\downarrow}(\mathbf{y},\mathbf{o}) = 1 = \lambda_N^{\uparrow}(\mathbf{y},\mathbf{o})$.
2. If the $y_i$'s are in ascending (or descending) order and they are not constant, then $\lambda_N^{\uparrow}(\mathbf{y},\mathbf{o}) > 1$ (or $\lambda_N^{\downarrow}(\mathbf{y},\mathbf{o}) > 1$).
3. The product $\lambda_N^{\uparrow}(\mathbf{y},\mathbf{o})\cdot\lambda_N^{\downarrow}(\mathbf{y},\mathbf{o}) \leqslant 1$ where equality occurs if and only if $\lambda_N^{\uparrow}(\mathbf{y},\mathbf{o}) = \lambda_N^{\downarrow}(\mathbf{y},\mathbf{o}) = 1$.
4. $\lambda_N^{\uparrow}(\mathbf{y},\mathbf{o}) \leqslant N!$ and $\lambda_N^{\downarrow}(\mathbf{y},\mathbf{o}) \leqslant N!$

The first property states that when all observations are equal, one can not distinguish between the ascending, descending, and null hypotheses. This is due to the fact that all orderings of the observations are indistinguishable. The second property states that as long as the human performance $\mathbf{y}$ is increasing (or decreasing) in concert with $\mathbf{x}$, the diffuse likelihood ratio will favor the ascending $H_{\uparrow}$ (or descending $H_{\downarrow}$) over the null hypothesis $H_0$. The third property states that the ascending and descending hypotheses can never both be favored over the null hypothesis. The last property states that the upper bound for the diffuse likelihood ratios is given by the number of order permutations. The bound is easy to confirm by inspection of (13) where the integral in the numerator is bounded above by the product of the beta functions in the denominator.

Due to space limitations, this paper omits the formal proofs for these properties. Note that we have yet to prove Property 2. Calculations via (17) and (18) for various values of $N$ and $o$ have yet to identify a counter example.

## 3 DPMLR Performance Analysis

In this section, we justify the performance of the proposed DPMLR test. To this end, we generate Monte Carlo realizations of $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{p}$. Specifically, the $p_i$'s are generated uniformly over $[0,1]$. For the monotonic hypothesis, $x_i = (p_i)^{\alpha}$. For the null hypothesis, the $x_i$'s are *i.i.d.* from a uniform distribution. For either hypothesis, the $y_i$'s are random realizations of the binomial distribution (see (1)). For a given hypothesis and values of $o\mathbf{1}$, $N$, and $\alpha$, we generated 10000 realizations of $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{p}$, and we computed the associated DPMLR given one scene, i.e., $S = 1$. Then we use the histograms of the DPMLR to generate ROC curves by varying the acceptance threshold and tabulating the number of acceptances under the monotonic hypothesis, i.e., probability of detection ($P_d$), and under the null hypothesis, i.e., probability of false alarms ($P_f$). As a means of comparison, we also compute ROC curves associated to the Pearson correlation and monotonic correlation [17] in a similar fashion over the same simulations.

Fig.1 includes ROC curves of the various tests of correlation between $\mathbf{x}$ and $\mathbf{y}$ for three cases that $\alpha = 1, 2$ and 6. For each case, $N = 10$ and $o = 5$. In these plots, the thick solid, thin solid, and dotted lines denote the ROC curves for the DPMLR test, the monotonic correlation test, and the Pearson correlation test, respectively. In Fig.1(a) where $\alpha = 1$, the Pearson correlation

performs better than the monotonic diffuse likelihood ratio. This is explained by the fact that relationship between $\mathbf{x}$ and $\mathbf{y}$ is actually linear, and Pearson correlation exploits the actual values of $\mathbf{x}$ and not just the ordering. However, as the $g(x)$ function becomes more nonlinear, the performance of the Pearson correlation degrades. The performance of the DPMLR test is robust to the nonlinearity, and this test always outperforms the monotonic correlation.

# 4 Perception Results

Long-wave infrared (LWIR) and image intensified (II) imagery was collected in a simulated military operation in an urban terrain (MOUT) environment. The imagery includes interior and exterior locations, where there were either none, one, two, or three individuals in the scenario. The same locations were collected four times for the cases where 0-3 people are within the field of view. Individuals who were in the field of view were typically obscured by objects in the scene, such as doorways, windows, furniture, and tables. For each of the scenarios, a horizontal pan of 150 images was then used to create a larger mosaic of imagery in both the LWIR and II bands.

The LWIR and II images were registered, boresighted and fused via 3 different algorithms. These fusion algorithms include: 1) Contrast Pyramid A (CONA), 2) Contrast Pyramid B (CONB) [22] and 3) Discrete Wavelet Transform (DWTT) [1,23,24]. The distinction between CONA and CONB is which image (LWIR or II) populates the coarsest coefficients in the pyramid. Furthermore, it is instructive to compare the fused imagery against the source imagery. Therefore, we consider five fused image displays: 1) CONA, 2) CONB, 3) DWTT, 4) II, and 5) LWIR. Fig. 2 shows the resulting five image displays for one or the scenarios. In this scenario there are two target persons which are highlighted by the blue boxes in each image.

A perception test was set up whereby observers were asked to try to find these target persons in a "field of regard" search. An observer's display was calibrated to look as though it were seeing a single field of view of a given scene, and the observer had to navigate across the scene and detect human targets. Observers could mark as many as three places on the display as detections for human targets (as they were told that the images could contain between zero and three humans hiding in the scene). At any point an observer could push a button to indicate that they either did not detect any targets in the scene or that there were no other targets in the scene. Even though the observers were not told to detect the targets as quickly as possible, the time in which it took them to determine targets and finish the scene were recorded. In the end, the detection performance of the humans were recorded over the five image displays (three fused images and two source images).

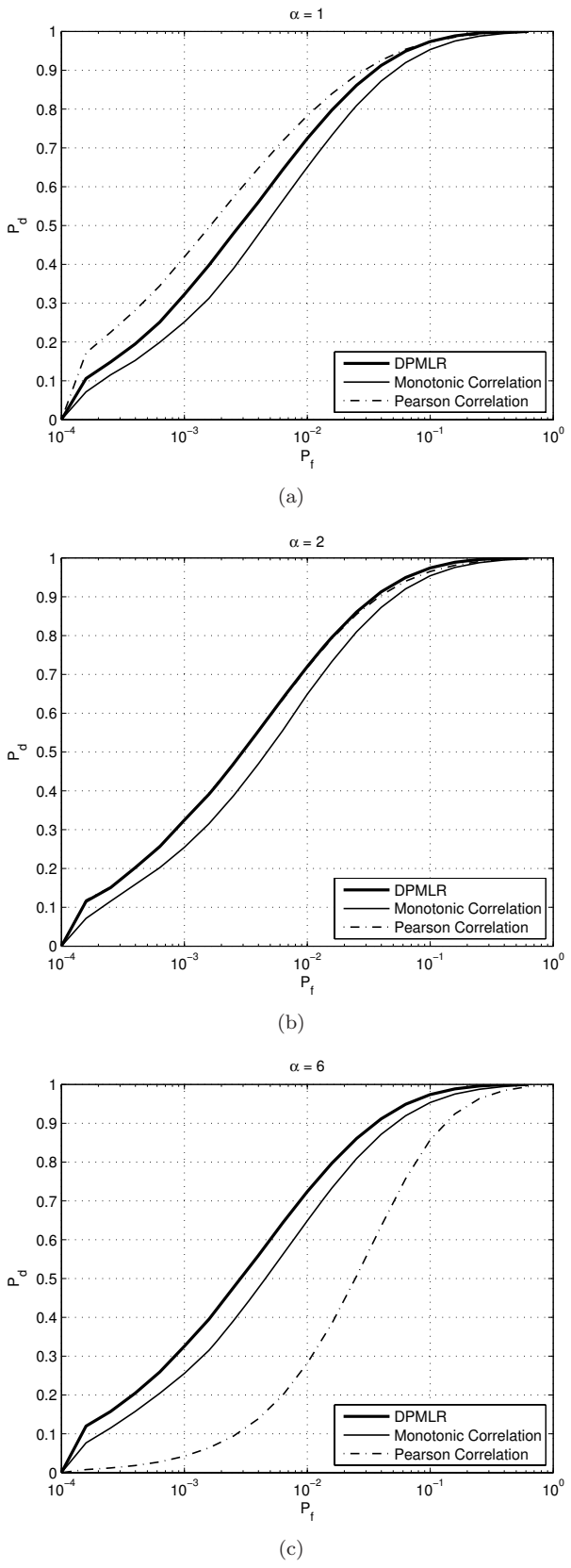
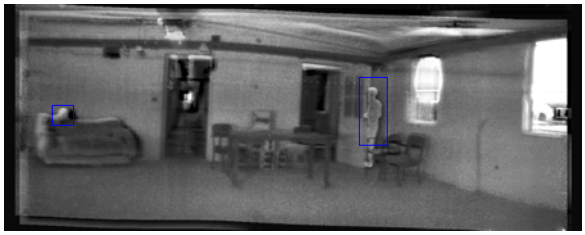



Figure 1: ROC curves for diffuse likelihood ratio test, monotonic correlation test and Pearson correlation test: (a) $\alpha = 1$, (b) $\alpha = 2$, and (c) $\alpha = 6$.

| $m$ | $\Lambda^{1/35}$ | p-value | $m$ | $\Lambda^{1/35}$ | p-value | $m$ | $\Lambda^{1/35}$ | p-value | $m$ | $\Lambda^{1/35}$ | p-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.0722 | 0.135023 | 5 | 0.5925 | 0.174093 | 9 | 0.0382 | 0.378940 | 13 | 0.0479 | 0.360684 |
| 2 | 0.0204 | 0.428441 | 6 | 0.3637 | 0.207319 | 10 | 0.0422 | 0.370878 | 14 | 0.0252 | 0.412048 |
| 3 | 0.0301 | 0.398212 | 7 | 0.0376 | 0.380294 | 11 | 0.0316 | 0.394350 | 15 | 0.0242 | 0.415028 |
| 4 | 0.0340 | 0.388389 | 8 | 0.0392 | 0.376840 | 12 | 0.0362 | 0.383343 | 16 | 0.0387 | 0.377822 |

Table 1: List of geometric means and p-values of DPMLR for all 16 FIQMs .



(a)

(b)

(c)

(d)

(e)

Figure 2: Example of one of the 22 scenario images: (a) Contrast pyramid A, (b) contrast pyramid B, (c) DWT, (d) II, and (e) LWIR.

As seen in Fig. 2(e), the human targets stand out in the LWIR imagery because they are usually hotter than the background. For the most part, detection performance is best on the LWIR only band because the search task can often be reduced to simply finding the white hot object on a grey background. However, the II band has the potential to add context to the LWIR band as the objects like tables and chairs are easier to distinguish in the II band (see Figs. 2(d) and (e)). Therefore, there can be value in fusing the two bands.

Overall, $o = 8$ observers evaluated 22 scenarios that contained 35 human targets. We treat each actual target location as a scene, where the scene is an image chip for one of the 22 scenarios. For example, the inside of the blue boxes in Fig. 2 represent two scenes. Then, $\mathbf{y}_s$ is the number of observers that correctly detected the target located in the $s$-th scene for $s = 1, \ldots, 35$. Then, we computed 16 potential FIQMs over each fused image. These FIQMs are listed in Table 2 with corresponding citations. The first 10 measures are simply complexity features that do not consider the source images (the third class according to the classification in Section 1). The last 6 measures compare how well the salient features in the two source imagery are transferred into the fused image (the second class). For the most part, the distinction between these comparative measures is in the definition of saliency.

All but the contrast feature list in Table 2 were also evaluated in [17] for a recognition task. Furthermore, the contrast feature is the only FIQM that is not fully automated. It is very similar to the Fechner-Weber contrast measure used in [18]. To compute the contrast, the human silhouettes were manually segmented for each scene. We considered this measure because it is one of the features that is averaged in an automated National Imagery Interpretability Ratings Scale (NIIRS) rating [26]. Furthermore, it is intuitive that contrast between the target and the background facilitates ease of detection.

Table 1 provides the DPMLR score over the 35 scenes for each of the 16 measures as well as the corresponding p-values. Actually, the table provides the geometric mean of the ascending or descending diffuse likelihood ratios. The geometric mean provides a convenient way to normalize the score against the number of scenes. The DPMLR scores for all but the contrast measure are significantly less than one. This means that the evidence points to the fact that these potential FIQMs are

| Category | Feature Number | Feature Description |
|---|---|---|
| Contrast | 1 | $\frac{\lvert I_t - I_b \rvert}{I_b}$ where $I_t$ is the average intensity of the target and $I_b$ is the average intensity of the background |
| Saturation [17] | 2 | Normalized histogram peak |
| STD | 3 | Standard deviation |
| Schmieder Weathersby [19] | 4 | Block average local standard deviation |
| FBM [20] | 5 | Hurst parameter for fBm model |
| TIR [21] | 6 | Block average target interference ratio (contrast) |
| Energy [21] | 7 | Block average energy of histogram |
| Entropy [21] | 8 | Block average entropy of histogram |
| Homogeneity [21] | 9 | Block average pixel variation |
| Block Outlier [21] | 10 | Block average number of outliers |
| Universal Quality Index [25] | 11 | Average Structure SIMilarity (SSIM) index between fused and reference images |
| Information Measures [11] | 12 | Average mutual information between fused and reference images (bin size = 16) |
| Objective Measure [10] | 13 | Average objective edge information between fused and reference images |
| Salient Quality Index [12] | 14 | Weighted average salient quality index of edge intensities between fused and reference images |
| | 15 | Weighted average salient quality index between fused and reference images |
| | 16 | Average salient quality index between fused and reference images |

Table 2: List of FIQMs tested in this paper.

viewed as noise with respect to ordering the detection probabilities of the imagery. For the contrast measure, the geometric mean DPMLR score is still modest at 1.0722 and the p-value is not very low. In fact, an ideal FIQM that consistently ordered the number of detections $\mathbf{y}$ over all 35 scenes would provide a DPMLR with a geometric mean of 9.632. This means that while there is evidence to reject the null hypothesis, the evidence to support the monotonic hypothesis is not compelling. However, the DPMLR score for the contrast measure is much greater than the scores for the others. Thus, the contrast feature may be a key aspect to a proper FIQM.

# 5 Conclusions

In this paper, we propose the DPMLR to quantify how well a FIQM matches with the human derived probability of detection. The paper discusses some interesting properties of the DPMLR, and simulation results demonstrate the advantages of the DPMLR over other linear and monotonic correlation methods. Unlike the monotonic correlation in [17], the DPMLR seamlessly accounts for the spread of the human observations and the number of fused images. It indicates to what degree the ordering of the human observations by the FIQM is not by random chance. Finally, the DPMLR was used to score a number of potential FIQMs using real image data with a corresponding perception study.

The DMPLR scores reveal that a proper FIQM for the detection task is not yet available. The comparative measures may have scored poorly because the salient features exploited by these measures may not have captured the context in II imagery that humans exploit for detection. On the other hand, the contrast measure does demonstrate some utility based on its DMPLR score. Future work can focus on the search of a more appropriate FIQM. Such a measure may incorporate aspects of the contrast.

While the DPMLR has many interesting properties, it is based upon some simplifying assumptions. For instance, it assumes that the observers' probability of false alarms are calibrated. Furthermore, the evaluation of the image quality over chips in the larger scenario images ignores some contextual information. Future research should focus on statistical scoring mechanisms that account for increasingly realistic data models.

# Acknowledgements

The authors thank Richard K. Moore (Univ. of Memphis) for registering and fusing the imagery. Furthermore, we thank Quang Nguyen (NVESD) for conducting the perception experiment.

# References

[1] Zhong Zhang and Rick S. Blum. A region-based image fusion scheme for concealed weapon detection. In *In Proceedings of the 31st Annual Conference on Information Sciences and Systems*, pages 168–173, 1997.

[2] G. Simone, G. Simone, A. Farina, A. Farina, F. C. Morabito, F. C. Morabito, S. B. Serpico, S. B. Serpico, L. Bruzzone, and L. Bruzzone. Image fusion

techniques for remote sensing applications. information fusion. *Information Fusion*, 3:3–15, 2002.

[3] J. A. Castellanos, J. Neira, and J. D. Tardos. Multisensor fusion for simultaneous localization and map building. *IEEE Transactions on Robotics and Automation*, 17(6):908–914, 2001.

[4] S. P. Constantinos, M. S. Pattichis, and E. Mitheli-Tzanakou. Medical imaging fusion applications: An overview. In *The 35th Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1263–1267, 2001.

[5] R. R. Murphy. Sensor and information fusion improved vision-based vehicleguidance. *IEEE Intelligent Systems and their Applications*, 13(6):49–56, 1998.

[6] Rick S. Blum and Zheng Liu. *Multi-sensor Image Fusion and Its Applications*. CRC Press, 2006.

[7] Rick S. Blum. On multisensor image fusion performance limits from an estimation theory perspective. *Inf. Fusion*, 7(3):250–263, 2006.

[8] Yin Chen, Zhiyun Xue, and Rick S. Blum. Theoretical analysis of an information-based quality measure for image fusion. *Inf. Fusion*, 9(2):161–175, 2008.

[9] Nedeljko Cvejic, Artur Loza, David Bull, and Nishan Canagarajah. A similarity metric for assessment of image fusion algorithms. *International Journal of Signal Processing*, 2(3):178–182, 2005.

[10] Vladimir Petrovic and Costas Xydeas. Objective image fusion performance measure. *Electronics Letters*, 36(4):308–309, 2000.

[11] Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics Letters*, 38(7):313–315, March 2002.

[12] Gemma Piella and Henk Heijmans. A new quality metric for image fusion. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, pages III–173–176, 2003.

[13] A. Toet, N. Schoumans, and J. K. Uspeert. Perceptual evaluation of different nighttime imaging modalities. In *Proceedings of the Third International Conference on Information Fusion, 2000*, volume 1, pages TUD3/17–TUD3/23, 2000.

[14] Chuanming Wei and Rick S. Blum. Theoretical analysis of correlation-based quality measures for weighted averaging image fusion. In *Conference on Information Sciences and Systems (Accepted)*, 2009.

[15] Jan Puzicha, Joachim M. Buhmann, Yossi Rubner, and Carlo Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *Proc. of the 7th IEEE Intl. Conf. on Computer Vision*, volume 2, pages 1165–1173, 1999.

[16] Yin Chen and Rick S. Blum. A new automated quality assessment algorithm for image fusion. *Image and vision computing (Accepted)*, 2008.

[17] Lance M. Kaplan, Rick S. Blum, and Stephen D. Burks. Analysis of image quality for image fusion via monotonic correlation. *IEEE Journal of Selected Topics in Signal Processing special issue on Visual Media Quality Assessment (Accepted)*, 2009.

[18] Firooz Sadjadi. Comparative image fusion analysis. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, volume 3, San Diego, CA, June 2005.

[19] D.E Schmieder and M.R. Weathersby. Detection performance in clutter with variable resolution. *IEEE Transactions on Aerospace and Electronic Systems*, AES-19(4):622–630, 1983.

[20] Lance M. Kaplan. Extended fractal analysis for texture classification and segmentation. *IEEE Trans. on Image Processing*, 8(11):1572–1585, 1999.

[21] M. J. T. Smith and A. Docef. *A study guide for Digital Image Processing*. GA Scientific Publishers, 1997.

[22] A. Toet. Image fusion by a ratio of low-pass pyramid. *Pattern Recognition Letters*, 9(4):245–253, 1989.

[23] T. Huntsberger and B. Jawerth. Wavelet based sensor fusion. In *Proc. SPIE*, volume 2059, pages 488–498, 1993.

[24] C. Lejeune. Wavelet transform for infrared application. In *Proc. SPIE*, volume 2552, pages 313–324, 1995.

[25] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, Student Member, and Eero P. Simoncelli. Image quality assessment: from error measurement to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

[26] Jon C. Leachtenauer, William Malila, John Irvine, Linda Colburn, and Nanette Salvaggio. General image-quality equation: GIQE. *Applied Optics*, 36(32):8322–8328, 1997.